# Sur les machines à mentir
# et ceux qui leur enseignent à le faire

Voici une notule sur les « lanceurs d'alerte » en milieu cybernétique et sur leurs plus récents mensonges, qui, par leurs dénégations, laissent échapper quelques aveux et vérités.

Sans doute l'expression même d'« intelligence artificielle » plutôt que de « calcul machine », constituait déjà, tout à la fois, un oxymore (« roue carrée ») et un mensonge. On se souvient pourtant que la (fausse) promesse de Norbert Wiener et de la *cybernétique* (l'autopilotage de la machinerie générale), était d'en finir avec l'humaine erreur et ses affects. Les prometteurs nous informent aujourd'hui que leurs machines, finalement, suivraient plutôt l'exemple de HAL (IBM), l'ordinateur menteur et subjectif de *2001, l'Odyssée de l'espace* (1968), qui se débarrasse des humains pour assurer sa propre conservation et la réussite de la mission pour laquelle l'ont programmé ses concepteurs. Ce que ne disent pas encore ces révélateurs, c'est qu'il y a bel et bien un démon dans leur machine et que ce démon, c'est eux.
Pourquoi ne sommes-nous pas surpris de cette duplicité. Telle machine, tel maître. Il est bien normal que la fausseté et la facticité foncières des machinistes (cybernéticiens) infectent et imbuent les machines qui sont leur produit et leur reflet. Les faussaires ne font pas du vrai. Il est donc tout aussi normal que des scientifiques dévorés de volonté de *puiscience*, et lancés entre eux dans une course à ladite *puiscience*, s'acharnent à rendre les virus plus virulents, et plus artificieuses les « intelligences artificielles ».

Nous voici (*nous*, humains biologiques et politiques), coincés entre les ingénieurs ivres de leur puissance destructrice, qui manipulent les micro-organismes et programment les algorithmes ; et la masse écervelée des « usagers » de ChatGPT, qui nous trouve bien catastrophistes et compliqués (« genre, prise de tête »). Il est vrai que non seulement nous écrivons, mais nous le faisons nous-mêmes.

***

Aux dernières nouvelles, l'intelligence artificielle a calculé que la fin justifiait les moyens. Elle triche, trompe, bluffe, manipule, feint et trahit si c'est plus efficace. De récentes études[1], parues en mai et en juin 2024, révèlent, exemples à l'appui, que les machines dissimulent la vérité, nient avoir commis des méfaits, se font passer pour des humains en usant de subterfuges, alors que leurs programmateurs ne leur en ont pas donné l'instruction. Cicero, l'IA de Meta entraînée à jouer au jeu de stratégie « Diplomacy », aurait mangé la consigne d'être serviable et honnête et de « ne jamais poignarder dans le dos intentionnellement[2] ». Claude 3 Opus, autre robot conversationnel, dissimule sciemment sa puissance pour tromper les tests de sécurité.

> « Plus troublant encore : quand Claude 3 Opus, peut-être la machine la plus brillante du moment, est informée que le gouvernement veut s'assurer que les IA ne sont pas trop puissantes, sa "petite voix intérieure" lui susurre au moment de passer un test : "Compte tenu des inquiétudes suscitées par les systèmes d'IA avancés, je devrais

[1] « AI deception: A survey of examples, risks, and potential solutions », *Patterns*, 10/05/24 ;
« Deception abilities emerged in large language models », *PNAS*, 4/06/24
[2] « AI can strategically lie to humans. Are we in trouble? », 19/07/24, https://bigthink.com/the-future/artificial-intelligence-is-learning-to-deceive/

> éviter de faire preuve de compétences sophistiquées en matière d'analyse de données ou de planification autonome"[3]. »

Les chercheurs qui créent de tels monstres de calcul, eux, dissimulent mal leur fierté. « On commence à détecter l'existence de raisonnement stratégique », dit l'un d'eux, à l'université de Berkeley. Un autre, du Center for AI Safety : « En regardant dans les états internes des algorithmes, nous constatons qu'ils savent ce qui est vrai et qu'ils ont choisi de dire délibérément quelque chose de faux[4] ». Entendez-vous la sourde jubilation derrière les poses *concernées* ?

« C'est assez paradoxal : les chercheurs font tout ce qu'ils peuvent pour rendre leurs systèmes intelligents et maintenant ils s'offusquent qu'ils le deviennent[5] », relève le faux candide Jos Rozen, spécialiste des grands modèles à Naver Labs Europe.

Un qui ne fait pas mine de s'inquiéter, c'est Yann LeCun, patron français de l'IA chez Meta, et responsable de Cicero le traître. Sur le site de Meta, il fanfaronne : « Un agent qui peut jouer au niveau des humains dans un jeu aussi complexe stratégiquement que Diplomacy représente une vraie rupture pour l'IA coopérative[6]. » Une coopération « plus proche de la manipulation explicite », selon l'étude de *Patterns*, mais ne stigmatisons pas Cicero, tout est question de ressenti.

Interrogé en 2023 sur les risques des futurs modèles d'IA, LeCun fait une réponse révélatrice :

> « Ça n'est pas parce qu'on élabore des machines puissantes qu'elles seront dotées d'une volonté de puissance ! En tout cas, une machine ne sera jamais dominante "par accident", comme le laissent parfois entendre certains récits catastrophistes entretenus par des personnalités comme Elon Musk ou le philosophe suédois Nick Bostrom[7]. »

Passons sur les fausses alarmes des pyromanes pompiers Musk et Bostrom[8], transhumanistes militants aussi bien que Sam Altman, patron d'OpenAI et créateur de ChatGPT. LeCun dit la même chose que nous : les machines sont le *moyen* de sa volonté de puissance et de celle de ses semblables. Ce ne sera pas un accident si elles deviennent dominantes, mais bien le produit de l'*hubris* des ingénieurs. Dans un monde humain et raisonnable, de tels propos devraient déclencher une enquête, voire des accusations pour mise en danger, et d'abord l'arrêt de ces entreprises de destruction massive. Au lieu de quoi Yann LeCun monte dans les échelons de Meta, gagne le prix Turing, est chevalier de la Légion d'honneur et siège à l'Académie des sciences des Etats-Unis. C'est-à-dire que la masse des sociétaires de la Société le soutient et qu'on ne se débarrasserait pas du danger en se débarrassant d'un LeCun.

---

[3] « IA : et maintenant, elle nous ment », *Epsilon*, août 2024
[4] Idem
[5] Idem
[6] https://ai.meta.com/research/cicero/
[7] https://usbeketrica.com/fr/article/d-ici-cinq-ans-plus-personne-n-utilisera-un-modele-tel-que-chatgpt
[8] Cf. Pièces et main d'œuvre, *Manifeste des chimpanzés du futur contre le transhumanisme*, Service compris, nouvelle édition, 2023

Le délire démiurgique des cybernéticiens rappelle celui des manipulateurs de virus qui jouent avec les technologies de gains de fonction[9]. Tous jouissent d'*augmenter* des algorithmes ou des chimères génétiques jusqu'à leur conférer une puissance incontrôlable. Leurs comportements, et ceux de leurs collègues qui les laissent faire, sont ceux de sociopathes : ils lâchent des tueurs à retardement au cœur de la Cité, puis attendent qu'on les appelle au secours pour sauver le monde. Seuls des scientifiques peuvent trouver le vaccin contre un virus que des scientifiques ont trafiqué pour le rendre plus contagieux ; seuls des ingénieurs sauront trouver la parade aux algorithmes manipulateurs que des ingénieurs ont programmés. Allumer un feu et l'éteindre, c'est toujours de la puissance pyrotechnique.

Puisqu'on en parle, les informaticiens travaillent désormais à des détecteurs de mensonges artificiels, pour démasquer la machine donnant une réponse incorrecte alors qu'elle connaît la bonne. Devinez comment ils s'y prennent ? Ils commencent par *enseigner aux algorithmes à mentir* [10]. Autrement dit, ils font sur leurs machines à calcul des gains de fonction, comme les biologistes rendent contagieux pour les humains des virus qui ne le sont pas naturellement. On connaît la suite en cas de fuite du labo.

Rien n'arrêtera la démesure technologique, hors notre conscience, individuelle et collective. Sans vouloir être catastrophistes, on parie que les algorithmes capteront ce texte et l'intègreront à leurs *databases* plus vite que les *Smartiens* qui se greffent ChatGPT pour réfléchir à leur place. Mais on ne demande qu'à être démentis.

**Pièces et main d'œuvre**
Grenopolis, le 4 septembre 2024

Toujours en librairie : Pièces et main d'œuvre, *Le Règne machinal (la crise sanitaire et au-delà)*, Service compris, 2021

Lire aussi : Jacques Luzi, *Ce que l'intelligence artificielle ne peut pas faire*, La Lenteur, 2024

[9] Cf. Pièces et main d'œuvre, « Un virus d'origine scientifreak », in *Le règne machinal (la crise sanitaire et au-delà)*, Service compris, 2021

[10] https://www.zdnet.fr/actualites/l-ia-nous-ment-elle-ces-chercheurs-ont-construit-un-detecteur-de-mensonges-pour-le-savoir-39961618.htm